**Babitha**

**E-Mail: Sudheer.r@tekprosol.com**

**Contact: 609 981 8255**

**SUMMARY**

Experienced Data Engineer with more than 12 years of extensive experience in designing, implementing, and managing data solutions. Specialised in AWS, Informatica, Databricks, Snowflake, Python, and PySpark. Demonstrated success in delivering high-quality data pipelines and driving business insights. Skilled in handling complex data integration challenges and optimising workflows for improved performance and scalability. Proficient in cloud-based platforms and big data technologies, Hadoop, and Spark, contributing to successful data-driven strategies. Exposure and knowledge on GenAI and LLM’s, AWS Bedrock, AWS CodeWhisperer, Azure Custom Vision and Azure Document Intelligence.

* Infrastructure design experience with GenAI congnitive tools: Azure Custom Vision and Document Intelligence.
* Hands on experience using AWS GenAI tools: AWS Q, Bedrock and CodeWhisperer
* Strong experience in Software Development Life Cycle (SDLC) including Requirements Analysis, Design Specification and Testing as per Cycle in both Waterfall and Agile methodologies.
* Strong experience in writing scripts using Python API, PySpark API and Spark API for analyzing the data.
* Extensively used Python Libraries in PySpark, Spark Performance Tuning
* Expertise in Python, Hive and Pig. Skilled in Power BI Desktop for data visualization, Reporting and Analysis.
* Developed reports, dashboards using power BI for quick reviews to be presented to Business and IT users.
* Extensive knowledge in Data warehousing concepts and ETL process. Experience in working with NOSQL databases like HBase and Cassandra. Experienced in creating shell scripts to push data loads from various sources onto the HDFS.
* Good Experience in implementing and orchestrating data pipelines using Oozie and Airflow.
* Hands on with Cloudera and Hortonworks distributions. Implemented Hive UDF’s
* Expertise working with Azure services ADF, Blob, Event Grid, Synapse, and AZ Copy.
* Hands-on experience with Amazon EC2, Amazon S3, Amazon RDS, VPC, IAM, Amazon Elastic Load Balancing, Auto Scaling, CloudWatch, SNS, SES, SQS, Lambda, EMR and other services of the AWS family.
* Hands on experience tuning Databricks spark Jobs
* Developed PySpark scripts from source system like Azure Event Hub to ingest data in reload, append, and merge mode into Delta tables in Databricks
* Optimized PySpark applications on Databricks, which yielded a significant amount of cost reduction
* Created Pipelines in ADF to copy Parquet files from ADLS Gen2 location to Azure Synapse Analytics Data Warehouse
* Environment: Azure ADF, Scala, PySpark, Spark, SQL, Snowflake, Databricks, GitHub, Azure Git, Kafka, ADF Gen2, ADF Blob Storage
* Hands on experience on Unified Data Analytics with Databricks, Databricks Workspace User Interface, Managing Databricks Notebooks, Delta Lake with Python, Delta Lake with Spark SQL.
* Good understanding of Spark Architecture with Databricks, Structured Streaming. Setting Up Microsoft Azure with Databricks, Databricks Workspace for Business Analytics, Manage Clusters in Databricks, Managing the Machine Learning Lifecycle.
* Can work parallelly in both GCP and Azure clouds coherently.
* Hands on experience in setting up workflow using Apache Airflow and oozie workflow engine for managing and scheduling Hadoop jobs.
* Solid understanding of snowflake.
* Used SNOWPIPE for continuous data ingestion from the S3 bucket.
* Developed Snowflake procedures for executing branching and looping.
* Created clone objects to maintain zero-copy cloning.
* Data validations have been done through information schema.
* Performed data quality issue analysis using Snow SQL by building analytical warehouses on Snowflake.
* Experience with AWS cloud services: EC2, S3, EMR, RDS, Athena, and Glue.
* Cloned production data for code modifications and testing.
* Performed troubleshooting analysis and resolution of critical issues.
* Hands on experience with snowflake utilities, Snow SQL, snow pipe, big data model techniques using python
* Experience in designing star schema, Snowflake schema for Data Warehouse, ODS architecture.
* Experience in Data Analysis, Data Profiling, Data Integration, Migration, Data governance and Metadata Management, Master Data Management and Configuration Management.
* Experienced in building Automation Regressing Scripts for validation of ETL process between multiple databases like Oracle, SQL Server, Hive, and Mongo DB using Python.

**EDUCATION**

* Master of Science in Computer Science, North Dakota State University, USA. (Dec, 2015)
* Master of Technology, Computer Science Engineering, Osmania University, India. (Oct, 2010)

**CERTIFICATION**

* Snowflake SNOWPRO Certification
* AWS Certified Data Analytics – Speciality – Feb - 2023
* Azure Data Fundamentals – Certification - Oct - 2022
* AWS Certified Solutions Architect – Associate – Jun - 2020
* Oracle PL/SQL Developer Certified Associate – Oct – 2016

**PROFESSIONAL EXPERIENCE**

**Capgemini**

**Role: Senior Data Engineer Feb 2022 – Till date**

**Responsibilities:**

* Monitoring the health checks of the ingested data pipelines from AWS, Palantir and other BP sources and sending notifications on the status of the pipelines, to the users with AWS SNS.
* Ingesting both batch and streaming data into medallion architecture.
* Real-time streaming data processing using Spark Structured Streaming and AWS Kinesis for continuous data analysis and insights.
* Bulk loading from the external stage (AWS S3), internal stage to Snowflake cloud using the COPY command.
* Loading data into Snowflake tables from the internal stage using SnowSQL and Snowsight.
* Used COPY, LIST, PUT, and GET commands for validating the internal stage files.
* Used import and export from the internal stage (Snowflake) from the external stage (AWS S3).
* Writing complex SnowSQL scripts in Snowflake cloud data warehouse for business analysis and reporting.
* Used FLATTEN table function to produce a lateral view of VARIANT, OBJECT, and ARRAY columns.
* Design and implement data transformation logic using PySpark, SQL on Databricks to clean, transform, and enrich raw data for analysis and reporting.
* Ingested data is sent to bronze data lake layer as raw data (json), moved to the silver data layer, where the data from the json file is converted to a delta table and in the gold data layer, using the STAR schema, fact and dimension tables created.
* Connecting to Athen from PowerBI to access the tables in gold data layer to show the analytics on the input pipelines.
* Interacted with end customers and gathered requirements for designing and developing common architecture for storing retail data within enterprise and building data lake in Azure cloud
* Developed Spark applications for data extraction, transformation, and aggregation from multiple systems and stored on Azure Data Lake Storage using Azure Databricks notebooks
* Written Unzip and decode functions using Spark with Python and parsing the XML files into Azure Blob Storage
* Build, scale and maintain data pipelines to process millions of daily transactions of customers from Redshift databases to AWS S3, Glue, accessible through Athena on an EMR cluster.
* Build datasets for capturing available customer journeys, third party deliveries, tokenisation, basket analysis and occasion analysis for both UK and Ireland markets.
* Provide support for deployed data applications and analytical models by being a trusted advisor to Data Scientists and other data consumers by identifying data problems and guiding issue resolution with partner Data Engineers and source data providers.
* Exploring new data sources to provide additional data to power data analytics team for Tableau visualisations.
* Code review and deployment in Code Commit.
* Data quality tests with Great Expectations framework and Glue DataQuality
* Internal and external stakeholder management.
* Mentor and assist fellow engineers both within and beyond my areas of responsibility and expertise.
* Workflow management using JIRA.
* Documentation captured in the confluence.

**Dixons Carphone – Multiplay ( UK ) Sep 2017 – Feb 2022**

Dixons Carphone is Europe’s number one Electrical and Telecommunications Retailer and Services Company, which operate across nine countries. Multiplay is a part of the business where the broadband, home phone and television services from top suppliers in the market are promoted to customers across the UK.

**Role: Data Manager**

**Responsibilities:**

* Managing pricing and promotional data for the Multiplay broadband products for various sales channels.
* Creating/adding new products/propositions in the pricing database
* Creating/updating services, equipment, packages (products) and offers (promotions) on the base Multiplay broadband packages/products.
* Coordinating with the trading team to decide the upcoming propositions.
* Launching products on various sales channels (Instore Retail, Indirect/Call centres, Ecommerce) on time and when it is needed.
* Creating workflows in CRM for gift cards, bill credits, cash back vouchers along with managing campaigns and partners in CRM.
* An expert in using the tool, the DMT (Data Management Tool), for managing the pricing database.
* Weekly reports on the sales and submission rates for individual broadband, phone and tv services providers on various sales channels: Ecommerce, Retail and Partnerships.
* Analysing the YoY and MoM sales data to identify new opportunities and the ways to reduce the customer orders pre submission and post submission dropout rate.
* Responsible for creating a monthly sales report showing the orders placed, cancelled, installed, and dropout rates from the sales cube.
* Identifying new opportunities and ways of working with the current system to improve the performance, efficiency, scalability, and usability of the existing pricing system.
* Active member in innovating new features and delegating the ideas to the developers and the stakeholders of the applications.
* Mentoring new team members in process, data, and applications.
* Documenting, updating, and managing the concepts and terminology used within the current pricing system.

**Microsoft, M&O CMO (Fargo, USA.) Jan 2016 – Jun 2016**

Microsoft is an American multinational company, which develops, manufactures, licenses, supports and sells computer software, personal computers and services. The project goal is to transform and consolidate marketing data from multiple source systems to one centralized place, which involved a lot of business and technical analysis, analysing and reporting on sales data across product categories and channels to support company sales and marketing efforts.

**Role: Data Analyst**

**Responsibilities:**

* Understanding, analysing, and transforming the data (marketing leads) from various marketing operations.
* Developed a wide range of Business Intelligence solutions used to manage output of large data volumes for marketing managers. SSIS packages developed and deployed using Team Foundation Server.
* Use ETL processes to achieve effective business use of high volumes of data.
* Build and maintain SQL scripts and complex queries for data analysis and extraction.
* Assisting with the transition to a new marketing automation system, analysing existing data practices and making recommendations for improvements.
* Generating reports using Power BI for various marketing teams.
* Ensure best practices are applied and integrity of data is maintained through security, documentation, and change management.
* Mentor and coach new team members on business process and technical aspects.

**Open Access Technology International (OATI) Minneapolis, USA Jan 2015 – Aug 2015**

Open Access Technology International, Inc. is the industry leader in the North American Energy Industry, providing advanced application solutions as web-based services to over 700 customers. The project that I worked on was aimed at Design, Development and Migration of OATI products from heterogeneous data sources, upgradation of Informatica and Oracle Database on Linux. The project involves Informatica data services to profile and document the structure and quality of all data.

**Role: Software Developer**

**Responsibilities:**

* Gathering business requirements for the data-warehouse as well as business-intelligence reports to be used by the management.
* Creating, monitoring, modifying, & communicating the project plan with other team members. • Worked on power centre tools including Designer/Repository, Workflow Manager/Monitor.
* Extensively used Informatica Transformations like Source Qualifier, Rank, Expression, Router, Filter, Lookup, Joiner, Aggregator, Normalizer, Sorter etc. and all transformation properties using Informatica Power centre 9.5.
* Translate Business processes into Informatica mappings for building Data marts by using Informatica.
* Conversion of WML (web plus) scripts into C# .net programs using .net framework.
* Responsible for writing test cases to cover overall quality assurance using HP ALM.
* Responsible for Integration, Functional and End to End testing.
* Involvement in Test Execution, Results Analysing and Defect Reporting.
* Attend daily status call with internal team and weekly calls with client and updated the status report.

**S&P Global Services June 2009 – May 2013**

**Role: Informatica Developer**

**Responsibilities:**

* Involved in complete life cycle of developing enterprise data warehouse application and developed ETL architecture using Informatica.
* Designed data warehouse target tables by using dimensional modelling techniques: Star and Snowflake schemas.
* Fact and dimension tables creation.
* Extract data from various data sources, transformed according to business requirements and loading into the targets.
* Created complex mappings using various transformations: Source Qualifier, Joiner, Aggregator, Expression, Filter, Router, Lookup etc.,
* Unit testing and user acceptance testing to check whether the data loads into target are accurate and satisfying the user requirements.
* Documenting in the confluence pages, mentoring and knowledge transfer